CA Journal of Cultural Analytics

Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature

05.23.16

Peer-Reviewed By: Hoyt Long

Clusters: Genre

Journal ISSN: 2371-4549 DOIs: 10.22148/16.003 (CA) and 10.7910/DVN/4ZVSKA (Dataverse)

In late Ming and early Qing China (1550-1700), unofficial historical narratives were extremely popular.[1] At the time, a variety of historical and semi-historical genres flourished, including *yeshi* 野史 unofficial or wild histories, novels on historical events, and dramas on historical events. [2] These texts form a group of related yet distinct styles of historical writing that I collectively refer to as quasi-histories. This is an umbrella term for a system of texts that contain some historical content but exist in a wide array of genres with different stylistic features and cultural significances. [3]

Yeshi and other quasi-historical texts, in spite of their late Imperial popularity, remain poorly understood as a result of a lack of attention from literary scholars, primarily because many were considered poorly written and of marginal literary value. This inattention is hardly surprising, given the time and energy necessary to properly analyze these texts using close reading methods. Despite this, these texts are important to understanding late Imperial literary and cultural production. *Yeshi* played an important role in information transfer and influenced how contemporary readers understood recent events, as did quasi-historical novels and dramas. [4] That they were such a popular genre of text, particularly among literati writing outside their official duties, further justifies additional scrutiny. An exploratory foray into quantifying the textual relationships among these works and more canonical texts allows me to integrate these works into my scholarship and elide normative judgments about quality. More importantly, it allows me to develop a stylistic taxonomy of quasi-historical works, which provides a basis with which to understand how they have been read and categorized.

Genre, as it relates to the quasi-historical texts discussed here, is a complicated concept. Genre is popularly understood as a grouping of texts constituted by a set of fairly uniform stylistic conventions. These conventions are amenable to quantitative analysis. Yet genre has a social dimension as well. Frederic Jameson has pointed out that genre is a socially constructed "contract" between reader and text, which defines expectations. [5] In the case of some genres, this is imposed from above by official bibliographers. In other cases, such as with drama, expectations developed organically. It is

[1]

[2] *Yeshi* form a very complicated genre of text. Lao Tie notes "*Yeshi* usually refers to privately authored historical books, in opposition to those produced under governmental authority (that is, "Official histories")." Lao Tie, ed, *Zhong hua ye shi ci dian* (Zhengzhou: Da xiang chu ban she, 1998), 2.

[3] These works primarily stand in opposition to "official histories" that were vetted by the Imperial government, which exerted strong control over the creation of historical writing. History in general was understood to be an undertaking meant to be "for officials, by officials." Endymion Wilkinson, *Chinese History: A Manual* (Cambridge: Harvard University Press, 2000), 20.3.

[4] David Wang has looked at how in the late Ming, novels on recent events constituted a "renewed concept of fiction as a way of registering the intelligibility of history." David Der-Wei Wang, *The Monster that is History: History, Violence, and Fictional Writing in Twentieth-century China* (Berkeley: University of California Press, 2004), 201. Han Li has also discussed these novels extensively in her research. For example, see Han Li, "News, History, and 'Fiction on Current Events': Novels on Suppressing the Chuang Rebellion" Ming Studies 2012.66 (2012), 56-75. This influence on historical imagination is a primary focus of my dissertation. It is noticeable in how representations of the Eunuch Wei Zhongxian 魏忠賢 (1568-1627) evolved in different genres of text after his death. Paul Vierthaler, "Quasi-histories and Public Knowledge: A Social History of Late Ming and Early Qing Unofficial Historical Narratives" (PhD dissertation, Yale University, 2014), chapter 1.

[5] Fredric Jameson, *The Political Unconscious: Narrative as a Socially Symbolic Act* (London: Routledge, 2002), 106.

sometimes impossible to disambiguate the social from the stylistic; many texts with similar content [6] but written in different genres circulated in distinct cultural strata. *Yeshi* and other quasi-historical texts exist in a space where the cultural construction of genre sometimes conflicts with its stylistic construction.

The content, stylistic, and generic nature of some of these historical writings inspired a fair amount of controversy and *yeshi* were the most problematic. Although *yeshi* never became an official bibliographic category, many were conventional historiographical works that mimic the style of official historical documents. [7] Others were impossible to distinguish from novels in both style and content. This dual nature led to many warnings by both contemporary and modern critics: *yeshi* are valuable because they lend insight into events ignored by official documents, but their salacious and often fictional nature make them inherently untrustworthy. [8]

In the late Imperial period, many *yeshi* were categorized by official bibliographers as *xiaoshuo* 小說 "petty talk." The philosophical underpinning of this classification emerged in *The Book of the Han* 漢書, a first-century history written by Ban Gu 班固. In his *Treatise on Literature*, Ban states that texts in this category relay that which is told on the streets, and were written by *baiguan* minor functionaries. [9] By the Ming dynasty, *xiaoshuo* had developed into a unique genre of literature with its own stylistic conventions and is most often translated as "novel." This development, and the persistence of bibliographers in classifying *yeshi* as *xiaoshuo*, led to confusion about the nature of *yeshi* and has caused dissatisfaction among both modern historians and late Imperial readers. [10]

While close readings and analysis seem to stylistically align *yeshi* with official historical writings (specifically 正史 *zhengshi*), it is sometimes difficult to evaluate their relationship with *xiaoshuo*, particularly in marginal cases. [11] Investigation into this relationship is further hampered by the number of *yeshi* published at the time, as well as the large number of novels. Hundreds, if not thousands, of *yeshi* were written during the Ming and Qing dynasties. Reading each work individually allows scholars to understand the style of individual works, but it is very difficult to develop a comprehensive and generalizable characterization.

Understanding the relationships among Chinese texts of different genres, and parsing what causes differentiation, is important. Can we understand the relative differences among novels, *yeshi*, and official histories with any amount of rigor? On initial examination, the archetypical examples of novels and official histories often appear distinct and stylistically unrelated. Yet there are historical and stylistic relationships between fictional prose and the biographical sections of *zhengshi*, as scholars such as Sheldon Lu have noted. [12] Quantitative analysis allows us to understand the broad middle ground between the two distinct poles formed by purely fictional and very historical writings, bringing apparently very different types of works into accord with each other by quantifying the nature of their differences.

This article uses statistical and linear algebraic analysis of term frequency lists calculated from digitized transcripts of quasi-historical texts to situate texts of various genres in relation with one another. Although a comprehensive sample of quasi-historical texts dating from the late Imperial period has not yet been digitized, the analysis that is possible using already digitized works allows me to effectively explore the differential stylistic nature of a number of quasi-historical genres. This analysis speaks to the combined influence that content and genre have on style.

I use stylometric analysis to visualize the relationships between documents in a manner that highlights the similarities and differences between fictional and historical texts. By quantitatively analyzing digitized texts, I can judge whether internal textual evidence suggests if *yeshi* themselves form a cohesive category, and visualize their relationship with other historical and semi-fictional narratives. It also offers insight into the extent to which historical content is predictive of style.

I find that *yeshi* stylistically mimic official historical works, are closely related to classical language texts, and often utilize more formal language structures. [13] In the first portion of this paper, I show that quantitative cluster analysis can clearly

[6] Content is *related* to genre, but only at a secondary level. Its influence is propagated through both style and the "social contract" Jameson discussed.

[7] *Zhonghua yeshi* 中華野史 (Jin'an: Taishan chubanshe, 2000), preface.

[8] Wang Shizhen (1526-1590) writes in his introduction to the *Shicheng Kaowu* 史乘考误 *Research into Errors in History* that one must weigh the benefits and dangers of relying on *yeshi* when understanding historical events. Wang Shizhen, *Yan shan tang bie ji* 弇山堂別集 juan 20, 1.

[9] Xie Guozhen 謝國楨, *Ming mo qing chu de xue feng* (Shanghai: Shanghai shu dian chu ban she, 2004), 82.

[10] Shao Yiping has criticized the use of the *xiaoshuo* category as a "trash can" for difficult-to classify texts like *yeshi.* Shao Yiping and Zhou E, "Lun gu dian mu lu xue de 'xiao shuo' gai nian de fei wen ti xing zhi. Discussing the non-generic quality of classic bibliographic studies' conception of 'Novel,'" *Fudan xuebao she hui ke xue ban* 2008.3 (2008), 10.

[11] Zhengshi are official dynastic histories. They are not the only type of official historical writ-ing, but they offer a large, diverse corpus to analyze.

[12] Sheldon Lu, *From Historicity to Fictionality: The Chinese Poetics of Narrative* (Stanford: Stanford University Press, 1994), 7.

[13] Although it is outside the scope of this article, I speculate this was a way to acquire legitimacy.

distinguish *yeshi*, novels, and dramas. I extend this analysis in the second half of this paper to explore how lexical features contribute to these stylistic relationships. This analysis shows that official histories, *yeshi*, and novels fall along a continuous gradient that is defined along several dimensions by the variable use of certain key terms that are strongly correlated with classical versus vernacular uses of late Imperial Chinese. It is significant that there is a gradient, rather than a sharp distinction between the poles.

Corpora

I use three related corpora in this research. The first is a small corpus of fourteen texts, chosen based on their applicability to Ming and Qing literary studies generally and the role of many as quasi-history. I incorporate the four most famous Ming novels, the late Yuan/early Ming *Water Margin* 水滸傳, the *Romance of the Three Kingdoms*, the *Journey to the West* 西游記 and the *Plum in the Golden Vase: Cihua edition*. I also include four Ming and early Qing dramas, the *Records of the Pure and Loyal,* the *Peach Blossom Fan* 桃花扇, the *Tale of the Lute* 琵琶記, and the *Peony Pavilion* 牡丹亭. The former two are relevant because they are both quasi-histories focused on historical events in the late Ming and early Qing, while the latter two are famous examples of the genre. I also include the *National Fragrance* 國色天香, a classical language novel, and the *Vernacular Romance of a History of the Woodcutters*, a vernacular language quasi-historical novel. I also include four Ming dynasty *yeshi*, the *The Captured Wilds from the Wanli Reign*, the *Biography of a Ming Emperor* 皇明本記, *A Narration of an Emperor's Grand Celebration* 皇明盛事述, and *A Wild Account* 野記. These texts are not representative of quasi-history in a meaningful way. Instead, they represent a humble entry into the larger project of understanding these difficult works. This shallow dip into a pool of several quasi- and non-quasi-historical texts offers a first step toward evaluating the homologies among late Imperial texts to determine if content or genre is more predictive of style. [14]

Next, I expand to 126 works procured mostly from Project Gutenberg and other places online to demonstrate the broad tendency for *yeshi* to cluster together. [15] Finally, I use a large corpus of 524 novels, historical romances, *yeshi*, and official histories representing around 540,000 pages. [16] This final corpus contains enough texts and is sufficiently refined in genre to offer a robust glimpse into the stylistic relationships among these works. Though most of these texts were written during the Ming and Qing dynasties, several are from earlier and later periods. Despite being only a small portion of Imperial Chinese literary production, these works represent a good initial departure into full-text digital analysis. [17]

Methods

Most of the analysis in this article is based on using a vector space model [18] to represent documents. I analyze the relationships among these texts with hierarchical cluster analysis and principal component analysis. Many of these techniques

[14] All corpora used in this article are available through DataVerse: Paul Vierthaler, 2016, "Late Imperial Chinese Texts: The Corpus for Fiction and History: Polarity and Stylistic Gradience in Late Imperial Chinese Literature", http://dx.doi.org/10.7910/DVN/GDYFAG, Harvard Dataverse, V1

[15] "Free e-books - Project Gutenberg," accessed March 10, 2014, http://www.gutenberg.org. The *yeshi* used in this analysis were found on wenxian.fanren8.com. Because the website does not include the generic label *yeshi*, I use the index of the *Zhonghua yeshi* section of the *Zhongguo lishi quanji: Zhonghua yeshi* (Beijing: *Zhongguo biaozhun chubanshe*) as a guide. I also include historical works that include *yeshi* and related terms in their titles.

[16] This corpus of texts was also acquired from wenxian.fanren8.com, an online collection of imperial Chinese texts. The last figure includes another 508 texts that can tentatively be considered *yeshi* (more on this later).

[17] The primary barrier to large-scale textual analysis in imperial Chinese studies is access to quality data. A variety of digitized late Imperial texts are available online, but they are sometimes difficult to access and often contain typographical errors. Some high-quality full-text databases either do not contain large collections of late Imperial texts or do not allow the bulk downloading of entire works. It is also very difficult to ascertain the accuracy of the transcriptions, given the size of the corpora. There are also many misnamed files: A version of the late Ming play the *Peach Blossom Fan* in wide circulation on the internet is actually a Qing novel, rather than the original play. I have done some rough checking to ensure the texts I include here are accurate enough to begin full-text analysis. This includes random spot-checking against printed editions found in the Institute of Chinese Literature and Philosophy at the Academia Sinica. As the field moves toward a more open-source ethos, the problem of access to accurately digitized texts will likely be mitigated. In line with this, there are currently projects underway that aim to crowd source the creation of high-quality transcriptions of Chinese texts. http://tenthousandrooms.yale.edu/, accessed December 17, 2015. Personal correspondence, Tina Lu.

[18] Conceptually, a vector space model understands documents to be points (or vectors) within a space whose axes are defined by the vocabulary contained within the corpus. The location of the document within said space is determined by how often each word is used in the document. For examples of the utility of these models, see Peter Turney and Patrick Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research* 37 (2010): 141-188.

have been widely used in linguistic corpus analysis and authorship attribution studies. [19] These two exploratory methods offer a way to cluster documents based on their internal vocabulary. J.H. Ward developed a hierarchical clustering algorithm (HCA) that I use to create dendrograms of textual relationships based on similarity scores. Closely related texts which use similar vocabulary exist close to each other in "vocabulary space." Ward's algorithm clusters the most similar texts together on branches of the dendrogram, with bridges between clusters based on how great the differences in vocabulary are. Principal component analysis, borrowed from linear algebra, offers a complimentary approach. This technique operates on the variance within the dataset. It creates new, abstracted axes on which to project the texts, essentially condensing a high number of dimensions into two dimensions that can be plotted on a sheet of paper. [20] In using these two exploratory analysis techniques, I can visualize different ways in which these texts relate stylistically.

The core of this study is done using term frequency lists, which are generated by tokenizing the documents and counting token frequency. Tokens are most commonly words, but can be any piece of a text, including punctuation or individual letters. [21] Words (or *ci* 詞 in the context of Chinese textual analysis), however, are a surprisingly accurate measure of a text's style, even when they are deprived of syntax and context. [22] While word frequency lists are excellent tools, they do have their drawbacks. As Stefan Gries warns,"they presuppose that the linguist (and/or his computer program) has a definition of what a word is and that this definition is shared by other linguists (and their computer programs)." [23]

Converting documents into countable units is a significant problem in Chinese, as there is a lack of natural delimiters within the text. [24] This problem is compounded when working with unpunctuated pre-modern works. In some cases, even understanding where sentences start and stop is a matter of debate. This results in a language that is difficult, though not impossible, to accurately parse without directly reading. Fortunately, 1-grams (or *zi* characters) are often discrete units of meaning in classical and vernacular Chinese. 1-gram analysis forms the of most of my analysis. *N*-grams offer a valuable measure of style and have also shown significant promise in authorship attribution algorithms. [25] [26]
The ideal number of most common tokens to use must be carefully considered. Christof Schöch has found that there is an inflection point where, as the number of most frequent tokens increases, the cluster shifts from being mostly representative of authorship to more representative of genre. [27] For Schöch, who is looking at French texts, this inflection point is

[19] Frederick Mosteller and David Wallace, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers* (New York: Springer-Verlag, 1984) is considered a classic of this type of study. See also J.F. Burrows and D.H. Craig, "Lyrical Drama and the 'Turbid Mountebanks:' Styles of Dialogue in Romantic and Renaissance Tragedy," *Computers and the Humanities* 28 (1994): 63-86 and JNG Binongo and MWA Smith, "The Application of Principal Component Analysis to Stylometry," *Literary and Linguistic Computing* 14 (1999): 445-466.

[20] It reduces the dimensionality of the dataset from one thousand dimensions to two, in the cases where the vectors contain the 1,000 most common 1-grams, so I can visualize the data. It essentially reorients the axes of the data for an optimal view of the results. Binongo and Smith refer to the action of PCA as "a translation and rotation of the original axes to arrive at a new coordinate system whose axes represent successive orthogonal lines of best fit." Binongo, "Principal Component Analysis," 459.

[21] Mathew Jockers uses punctuation among the 44 tokens he uses to explore Shakespearean genres. Sarah Allison et al., "Quantitative Formalism: An Experiment," accessed March 24, 2014, 10.

[22] Ted Underwood, "Wordcounts are Amazing," *The Stone and the Shell*, February 20, 2013, accessed March 24, 2014.

[23] Stefan Gries, Quantitative Corpus Linguistics with R: A Practical Introduction (New York, NY: Routledge, 2009), 12.

[24] For more on this, please see the segmentation appendix.

[25] Fuchun Peng, et. al., "Language Independent Authorship Attribution using Character Level Language Models," http://www.aclweb.org/anthology/E03-1053, accessed December 11, 2015.

[26] Significant textual processing is necessary prior to analysis. For each document, I discarded all punctuation and unimportant artifacts. In some cases I broke each work into equal length *n*-gram segments to ensure comparisons occurred across equal length texts. I also discarded the remaining unequal length section. I initially found the technique of dividing the texts into uniform sections in Binongo and Smith's "Principal Component Analysis," in which they created a table by "dividing these plays into blocks of 5,000 words (irrespective of the actual act and scene divisions)…" Binongo, "Principal Component Analysis," 448. Burrows and Craig use a similar approach and divide the plays they are interested in into 4,000 word segments. This approach is particularly useful when trying to determine the authorship of suspect sections of a work (Burrows and Craig, 67). In doing so, I can also visualize how internal segments of texts relate to other works. The length of segment influences the result. The shorter the section of text, the more likely small similarities emerge, which are more likely to be content-driven. The longer the text, the more the overall style matters, and the smaller changes within the text are smoothed over. Longer sections tend to cluster better according to known labels.When a text was in traditional Chinese characters, I converted the texts into simplified characters.
I calculate an *n*-gram frequency list for each section of text to build a representation of each section in a vector. Each vector includes the token score for the most common tokens found across all fourteen texts. Each vector represents a point (or line) in *n* dimensional space where *n* is the number of unique variables. Hierarchical cluster analysis using the Ward method clusters the most closely related vectors (and hence texts) based on a similarity measure. This method clusters vectors with the aim of reducing the total amount of variance within the cluster and groups the most similar vectors together. J.H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association* 58.201 (1963): 236-244. For my initial evaluation, I measure the Euclidean distance between the most-common-token vectors for each section of text and then hierarchically cluster them. Given that each text is of equal length, I use Euclidean distance in most analyses rather than cosine similarity or some other measure. For a discussion of various distance calculation methods see Christof Schöch, "Beyond the Black Box, or: Understanding the Difference Between Various Statistical Distance Measures," *The Dragonfly's Gaze*, August 3, 2012, accessed March 24, 2014.

[27] Schöch explores the extent to which the length of most frequent word lists matter and found that when using a short list authorship is the main factor that causes clustering. Genre becomes a large factor in clustering once a surprisingly high threshold is crossed. Schöch, "Author or Genre?"

somewhere around 750 common words, although other scholars have found different numbers effective. [28] He found the authorship signal in most common token vectors to be very strong, even when more than the one thousand most common words in the corpus are analyzed. [29] Scholars must evaluate what works best on a case-by-case basis. The relationship between top frequency threshold and authorship or genre detection also likely depends on the language in which the text is written: in Chinese, genre plays a very significant role in style and usually obscures authorship signal.

Hierarchical Cluster Analysis: Beginnings

Hierarchical cluster analysis offers the first glimpse into how fourteen Ming and Qing texts stylistically relate to each other. In Figure 1, an unrooted dendrogram illustrates the results. It is immediately obvious that vectors taken from each text tend to cluster quite closely with others taken from the same work forming clades, or groups of closely related texts that fall on the same branch of the dendrogram. Texts of similar genre form the next highest order of clustering.

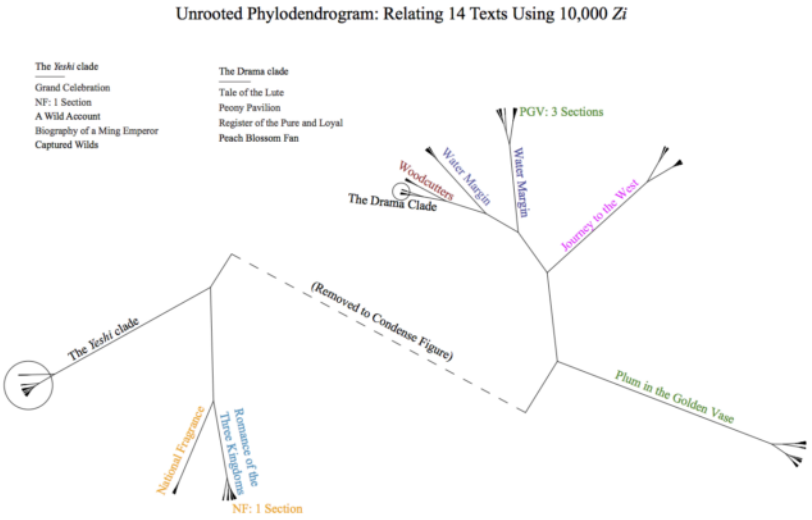


Figure 1. An Unrooted Phylodendrogram Showing 14 Texts using 10,000 *Zi* Segments and 100 Most Frequent *Zi*. The *yeshi* all fall on a single clade, as do the dramas. The *Water Margin* splits along two clades. Several sections of the *Plum in the Golden Vase* appear among the *Water Margins* sections. The texts with more classical language fall on the left side of the graph. Those on the right are more vernacular.

The relationship between *The Water Margin* and *The Plum in the Golden Vase* validates this approach as a measure of textual relationship. Three *Plum in the Golden Vase* vectors fall outside the *Plum in the Golden Vase* clade and align squarely within *The Water Margin* clade. This is compelling, as these vectors represent the first three 10,000 1-gram sections of the *Plum in the Golden Vase*. Most of this section of the *Plum in the Golden Vase* is taken almost verbatim from *The Water Margin*. Specifically, they come from the chapters describing the adventures of a man named 武松 Wu Song.

This figure also points to the divergence between novels and *yeshi.* On one side are classical language and predominately historical texts. All the *yeshi* cluster on this side, as well as the two novels the *National Fragrance* and *The Romance of the Three Kingdoms*. Further, all *yeshi* fall on this side and group on the same clade. The *Narration of an Emperor's Grand Event* is most distantly related to the other *yeshi*, the *Wanli yehuo bian* falls onto two small clades, and the other two yeshi fall nearby.

The proximity of the classical language *National Fragrance* near the *yeshi* suggests that the language found in *yeshi* is closely related to the language in classical novels. It is also not surprising that *The Romance of the Three Kingdoms* is so

[28] Matthew Jockers only needed 44 to categorize a number of English novelistic works. Allison, "Quantitative Formalism," 6.
[29] Schöch, "Author or Genre?"

closely related to these works. It is similar both in terms of content (being mainly focused on history), and its language, which is more formal than in some of the other novels. [30]

Vernacular works dominate the opposite side of this dendrogram. However, the works shown here are more distantly related to each other than those on the classical side of the figure; both the *Plum in the Golden Vase* and the *Journey to the West* exist relatively independently of other vernacular works. Still, they are much more distantly related to the *yeshi* than to other nearby vernacular works.

The plays the *Peach Blossom Fan*, the *Registers of the Pure and Loyal*, the *Tale of the Lute*, and the *Peony Pavilion* bear a similar a degree of homology as is found among the yeshi, despite the significant difference in content. This seems to suggest that genre plays a significant role in how these texts cluster. The close relationship between the plays and the *Woodcutter* is somewhat surprising.

It seems unlikely that much of the homology in this particular sample is derived from historical subject matter *per se*. Note that *The Woodcutters*, the *Peach Blossom Fan*, and the *Registers of the Pure and Loyal* are all works on historical topics from the late Ming and early Qing dynasties, yet fall quite far from the *yeshi* and the classical novels. Additionally, all plays cluster closely together, suggesting that historical versus non-historical is not a determining factor on this level of analysis. Most likely there is a hierarchy of influencing factors, starting with style of language, as these other works are more vernacular, followed by genre. The close proximity of these works on the dendrogram may suggest that a third level of clustering, based on content, may be occurring. The *Woodcutters* and the historical plays all contain similar historical content based on events during the Ming and Qing transition or the thirty years leading up to that. [31]

The four *yeshi* contain 1-gram frequencies that are closely related to those in the classical language works. How closely, then, do they relate to an official history? A second cluster analysis conducted to include the very long *Official History of the Ming Dynasty* answers this question. Figure 2 shows these results but omits the side of the dendrogram containing the vernacular works, which is not significantly different from Figure 1. [32] The *yeshi* are quite close to sections of the *Official Ming History*, almost suggesting that these *yeshi* were incorporated into the *Official Ming History* in some way. The vectors representing the *Official Ming History* are more dispersed through the dendrogram than the other works. This is possibly because it contains a wide range of topics, sections with different styles, and was written collaboratively by many government scholars. The novels stand slightly independently of the historical works (although the *National Fragrance* appears closely related to part of the *Captured Wilds*).

[30] Although *The Three Kingdoms* is often categorized as a vernacular novel, it contains many elements of classical Chinese. It is listed in the *bai hua* 白話 section of the *Zhongguo gudai xiaoshuo zongmu tiao* 中國古代小說總目提要 index. *Zhongguo gudai xiaoshuo zongmu tiyao* 中國古代小說總目提要 (Beijing: Renmin wenxue chubanshe, 2005), 32. Anne McLaren refers to its language as "simplified classical." Anne McLaren, "Constructing New Reading Publics in Late Ming China," in *Printing and Book Culture,* ed. Cynthia Brokaw, (Berkeley: University of California Press, 2005), 176. Later in this analysis, this novel will be classified as a specific genre of fiction called the "historical romance."

[31] There are further interesting things happening with the *Water Margin*, aside from its relationship with the *Plum in the Golden Vase*. It clearly falls onto two different clades. One side, which includes the section related to the *Plum in the Golden Vase*, most likely corresponds to the first section of the book, which is focused on how the heroes of the story all arrived at Liangshan marsh. The second, which falls on a distinct clade, discusses the adventures of the bandits as a group. There are several explanations for this, some more speculative than others. The most likely explanation is the sharp turn in narrative style from the individual anecdotes to campaign style activities. One could speculate that two separate authors wrote the two parts, and the second author was unable to completely emulate the original's style. This is in line with scholarship that disputes Shi Nai'an's authorship. This dispute centers on who Shi Nai'an was, if he was Luo Guanzhong, or if they both worked on the text. The editor of the *Indiana Companion to Traditional Chinese Literature* claims that "evidence excludes the possibility of a single author…" and rejects Shi Nai'an and Luo Guanzhong as the author. William Nienhauser, Jr., ed, *The Indiana Companion to Traditional Chinese Literature* (Bloomington: Indiana University Press, 1986), 712.

[32] I do not include the *Official History of the Ming* in the initial analysis largely because of its great length; when divided into sections it dominates the cluster analysis, obscuring how closely the *yeshi* cluster together.

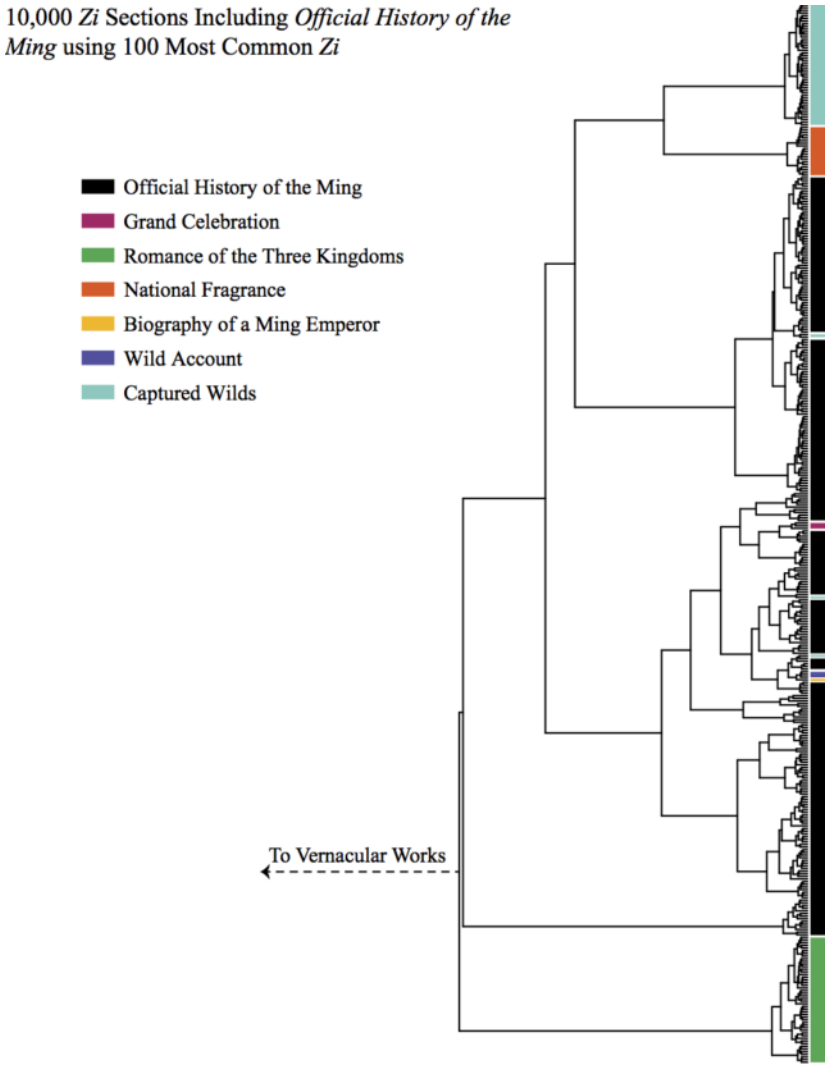


Figure 2. Phylodendrogram of 10,000 *Zi* sections Including *Official Ming History* using 100 Most Common *Zi*. There is a very close relationship between the *Official History and the* yeshi.

Hierarchical Cluster Analysis of 126 Complete Texts

To this point, I have used segmented sections of text as a rubric to judge the relationships among the internal components of these works. [33] Hierarchical cluster analysis provides a basis of comparison between these various works and highlights some interesting relationships. But to what extent can cluster analysis differentiate the generic features of a work, *yeshi* in this case? Are the features of *yeshi* homogenous enough to allow this to happen? Do *yeshi* still appear as cohesive as they are when looking at only four of them?

Rather than segment works into ten thousand token segments, what if we use the entire work, regardless of length? Doing so allows us to visualize a larger number of individual works without generating illegible visualizations; if segmented, the number of individual segments would reach into the thousands. Analyzing many works of different genres at once may provide more information on how *yeshi* relate to other works. The disadvantage of using whole texts is that the relationship between the internal components is no longer evident. It is also necessary to normalize the data to ensure long works do

[33] I borrow this technique from Burrows and Craig. Burrows, "Lyrical Drama," 67.

not overpower short ones. There are a number of ways to normalize the data, including by assigning each token a score based on the number of occurrences per ten thousand tokens. [34]

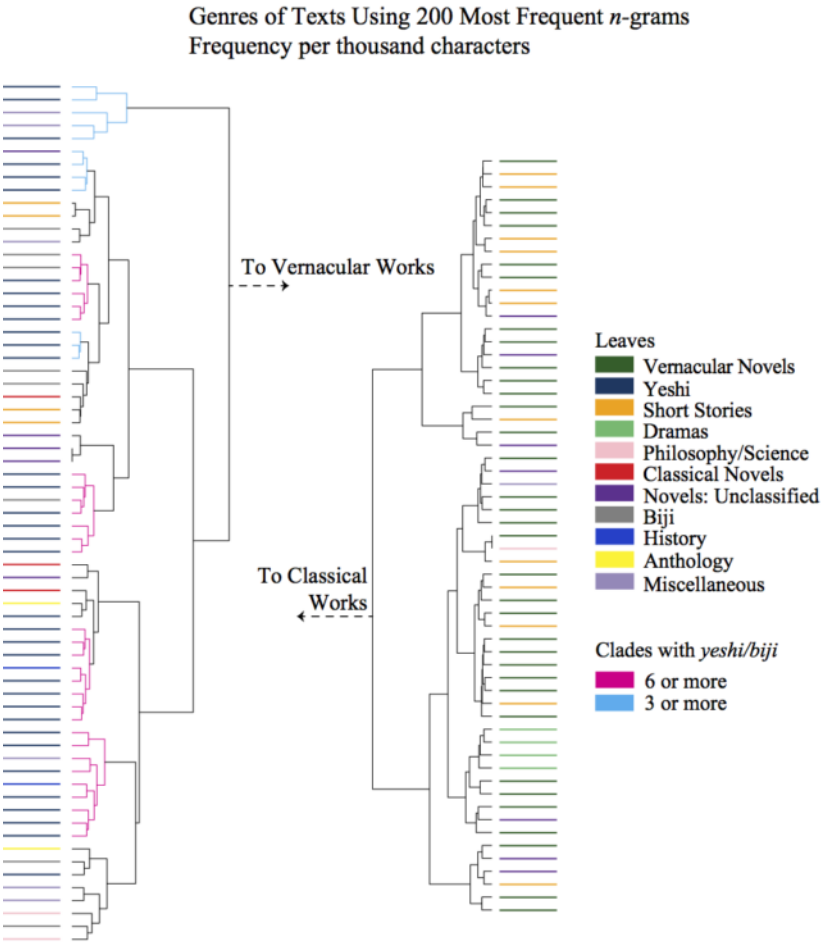


Figure 3. Hierarchical Clustering of Texts Using 200 most frequent *n* -grams, using their frequency per thousand characters, Vernacular novels all appear on the right side of the dendrogram, while classical works appear on the left. No *yeshi* appear on the right. Most of the *yeshi* cluster closely with other *yeshi* and are often only interspersed with official histories and *biji*. Clades dominated by these works are highlighted in magenta (for clades with at least six *yeshi* or *biji*) and cyan (for clades with least three). Note that both official histories cluster with *yeshi*.

Figure 3 is a rooted dendrogram of 126 different works produced using full-length texts. Calculating each n-gram's occurrence per thousand characters normalized the data. I use the two hundred most frequent tokens found in the corpus to cluster the works. [35]

Figure 3 clearly shows the large division between vernacular and classical works present in the other dendrograms. The *yeshi* form several very cohesive clusters only punctuated by a *biji*, several unclassified novels, and the two official histories. The vernacular novels show a similar tendency to cluster with each other, and the dramas remain tightly knit. The most widely distributed works are short stories and have segmented themselves based on the formality of language. Though the divisions are not perfect, the texts cluster along roughly generic lines.

These dendrograms allow us to explore quasi-history in new ways. Their grouping is consistently based on genre in a

[34] Ted Underwood describes another reliable method of normalizing the scores, derived from *chi*. He advocates calculating the distance between the number of actual occurrences of a word within a text and the number of occurrences expected if the text was representative of the corpus as a whole. "Tech Note" accessed March 25, 2014, http://tedunderwood.com/tech-notes/.

[35] All texts are longer than ten thousand tokens. Shorter texts were discarded. Eder suggests random sampling works well for authorship attribution in shorter texts, but here I simply excluded the short works. Eder, M. (2014), "Does size matter? Authorship attribution, short samples, big problem." *Literary and Linguistic Computing*, 29, advanced access (doi:10.1093/llc/fqt066).

manner that suggests a central difference between *yeshi* and other quasi-histories is the style of language. Further, it provides evidence that *yeshi* and vernacular novels are quite distinct. Here content does appear to affect the results. Some of these texts have almost identical content written in distinct linguistic registers: novels written about the eunuch Wei Zhongxian 魏忠贤 (1568-1627) fall on the right side (with the other vernacular novels) and the *yeshi* fall on the left (with the other *yeshi*).

Toward a deeper understanding of a fictional/historical stylistic gradient

It seems clear that *yeshi* cluster near classical language texts and away from vernacular novels. But what is the nature of this relationship (or lack thereof)? Are *yeshi* sensibly envisioned as threats to official history or as distinct from fictional prose? It is hard to draw broad conclusions with the small dataset previously used, so I now turn to a different, much larger corpus of Chinese texts to analyze the stylistic relationships among fictional, semi-fictional/semi-historical, and historical works in late Imperial China. In the following analysis, I look at 34 *zhengshi* (2,929 ten thousand token sections), 365 novels (5,631 sections), 57 *yeshi* (347 sections), and 68 *yanyi* historical romances (novels written on historical topics, 1,731 sections). [36] Most Imperial-era texts have between 140 and 280 characters per page, [37] which means a ten thousand token section represents between 35 and 70 pages of text. This results in a corpus of about half a million pages. With the exception of most of the *zhengshi* and half of the *yeshi*, the majority of these texts were written during the Ming and Qing dynasties. [38] I leave drama out of this analysis, because the stylistic conventions within these texts place them somewhat adjacent to the prose styles found in these other works. [39]

The cluster analysis shown in Figure 4 follows the same trends seen earlier. The broadest separation is still along classical/vernacular lines demonstrating a distinct classical to vernacular polarity. All official histories are on the left. The clade on the far left has almost all of the official histories and many *yeshi*. Some classical novels appear on the left of the dendrogram, clustered separately from the official histories. The historical romances complicate the picture. Some integrate into the novel clade on the right, while others sit amidst historiographic documents. The *yeshi*, which were written in primarily classical language, are most closely related to official histories, but show some affinity to the historical romances as well. [40] This mixing suggests that historical content leads fictional texts to cluster more closely with historical works. To a certain extent, historical content appears predictive of the type of language found in a work. Exploring the nature of this apparent mixing requires further study with principal component analysis; is it continuous bridging of style or a strict dichotomy? How exactly does historical content influence style?

Principal Component Analysis

Yeshi form a largely cohesive genre that clusters closely with official histories and with classical language works more generally. The close bibliographic relationship late Imperial bibliographers constructed between *yeshi* and *xiaoshuo* "petty talk" seems to be an outdated artifice, adding weight to the argument made by those who see a need to differentiate the two. Yet, late Imperial novelists' tendency to appropriate the term *yeshi* for their titles reinforces the connection with contemporary *xiaoshuo* "novels." At least two novels within the Wenxian dataset are classified as *yeshi* in some collections, while a variety of other novel include *yeshi* or related terms in their titles. [41] There is a clear difference between the

[36] These texts are all sourced from wenxian.fanren8.com, which makes many digital Chinese texts freely available. The *xiaoshuo* in this collection are "novels" as understood by the stylistic conventions of the late Ming and Qing dynasties, not the "petty talk" works from much earlier in Chinese history.

[37] Paul Vierthaler, "Analyzing Printing Trends in Late Imperial China Using Large Bibliometric Datasets", *Harvard Journal of Asiatic Studies* (forthcoming, 2016), Figure 6.

[38] The language within both *yeshi* and *zhengshi* is similar across time. Regardless, I later confine the analysis to just Ming and Qing works.

[39] Including dramas in the hierarchical cluster analysis shows unsurprising results. With several classical exceptions, they all cluster on a single clade on the vernacular side of the dendrogram. To perform the necessary calculations, I use the term frequency vectorizer implemented in the python package scikit-learn. This term frequency vectorizer normalizes the results using L2 normalization.

[40] When I insert an unrelated genre into the analysis, these texts tend to cluster among themselves, away from the historical and fictional types of works (even when they are primarily prose, such as poetic criticism). When running these tests, it was sometimes necessary to balance the input corpus so an equal number of sections come from each genre. Particularly when evaluating genres outside the ones considered here, as they have a very strong influence on the shape of the principal component space.

[41] These terms include outer history *waishi*, hidden history 逸史 *yishi* , and secret history *mishi* 秘史.

conception of the texts by bibliographers and the impressions they impart upon the average reader. Novel authors play with this dichotomy in interesting ways by incorporating *yeshi* into their titles. [42] Several cases of this are visible as *yeshi* clustering among vernacular novels in Figure 4.

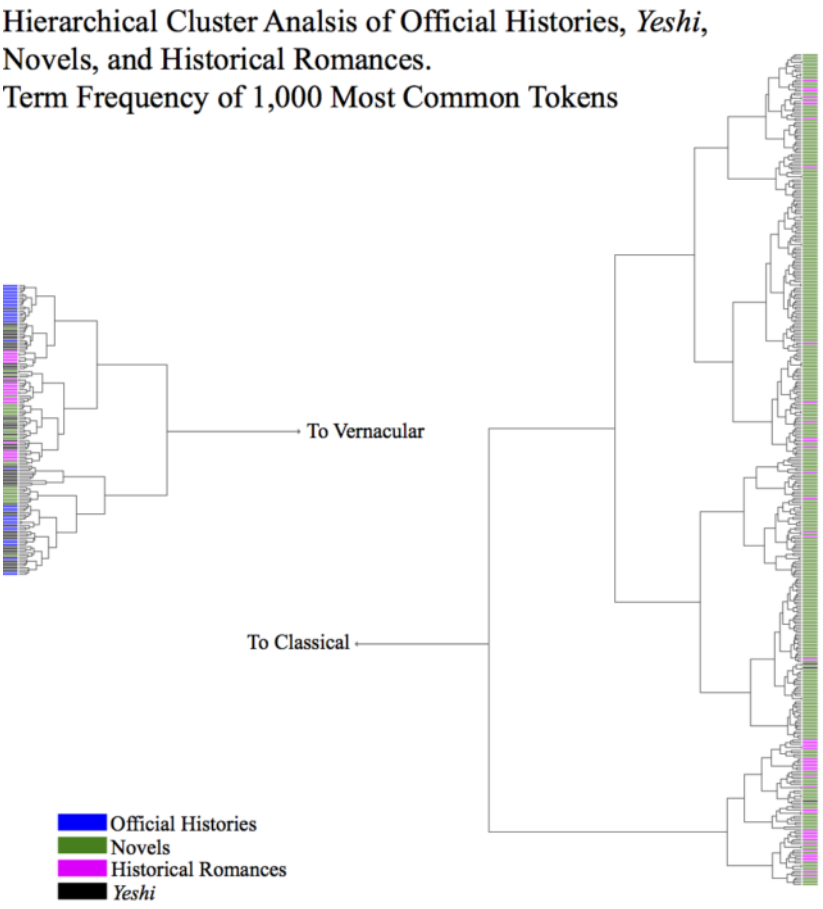


Figure 4. Hierarchical cluster analysis of official histories, *yeshi*, novels, and historical romances using euclidean distances of term frequencies of the 1,000 most common tokens, Texts are not divided into equal length segments. Like earlier examples, this cluster analysis divides very starkly along a classical/vernacular basis. Those few novels on the left side of the dendrogram are all classical language. On the right are the novels, with many historical romances and a few *yeshi*. Histories and unofficial histories cluster together, with only some interference with intervening historical romances and highly classical novels.

HCA offers insight into an unsurprising and large classical/vernacular divide in late Imperial texts. Identifying the nature of this divide in HCA, however, relies on *a priori* knowledge of a dataset's composition. Principal component analysis offers a clearer grasp of the relationships among these texts by parsing the variance within the dataset, exposing the exact vocabulary differences that constitute this divergence. Without *a priori* knowledge of the corpus, PCA sheds light on why the texts cluster as they do. It also allows me to visualize the texts in a new way; after calculating the principal components, the data can be graphed onto these components, which act as abstracted axes. Critically, this allows us to understand how much historical rhetoric influences the style of texts, and if this divergence is occurring because of some content-based semantic influence, or if historical discourse exerts influence on the very common words within the corpus.

To properly characterize the divergence between historical and fictional works, I start with a subset of the corpus that should show a clear divide: novels and official histories. Figure 5A shows ten thousand 1-gram sections from official histories and novels plotted onto the first two principal components. There is a clear distinction between the two genres of text, with only a few sections of novels intruding into the space occupied by the official histories. This backs up the

[42] An interesting avenue of further research would be to analyze both the contents of these novels and their historical context to better parse *why* they make this choice.

phenomenon seen in earlier figures. The clarity of the division, however, is somewhat surprising. The genres separate along a linear cleavage, with most distinction lying along the first principal component, which accounts for 33 percent of the variance within the dataset. The second principal component accounts for approximately 6 percent of the variance. [43]

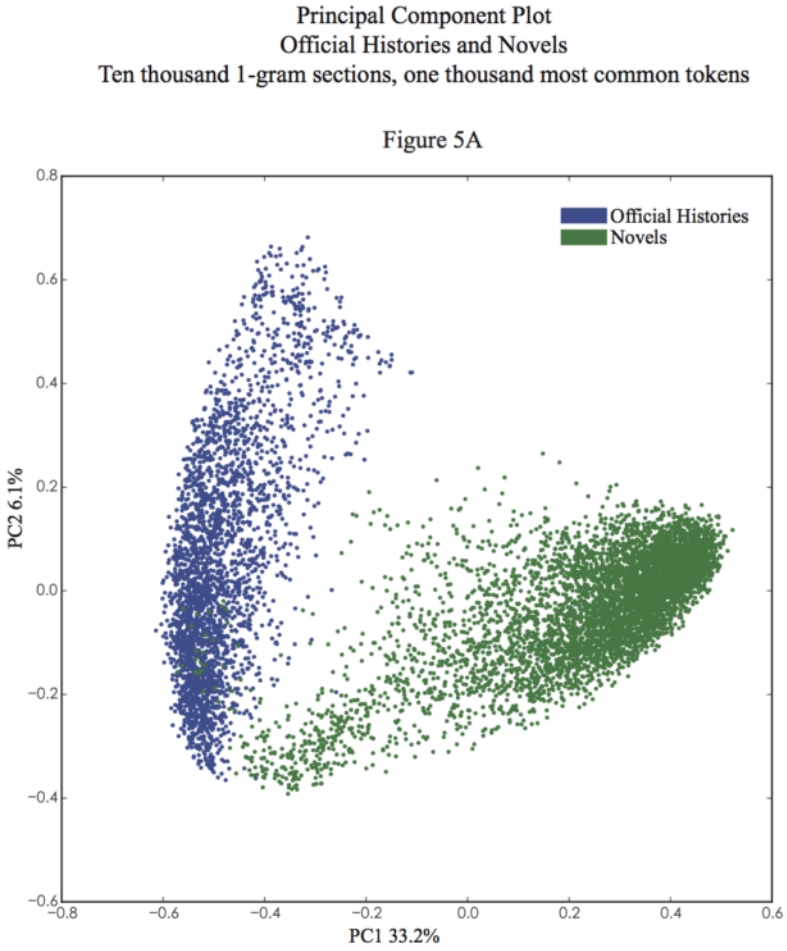


Figure 5A. Principal component analysis of official histories and novels in ten thousand 1-gram sections using the thousand most common tokens in the corpus. With only a few exceptions, sections of text from novels and official histories cluster when plotted along the first and second principal components. A few sections from novels do intrude into the space heavily dominated by histories. No histories enter into the dense novel space on the middle right of the PC space.

The clustering is evident, but to interpret this PCA appropriately, it is necessary to look at a loadings plot, which shows how each variable influences where documents fall in Figure 5A.

Figure 5B offers insight into the nature of the principal components. Several things are immediately clear. The first principal component is roughly correlated with the formality of the language found within the text. That is, with how classical or vernacular the text is. For example, 之 *zhi* and 的 *de* are both possessive particles and have roughly the same meaning, but the former is a classical particle while the latter is vernacular. [44] The second principal component is harder to parse but words related to time and geographical space stand out. They pull documents into the upper left quadrant (年 *year*, 月 *month*, 州 *province*, *south*, 西 *west*, 二 *two*, 三 *three*), a space that is formed entirely of official historical words.

This strong focus on temporality and space is unsurprising in histories. However, it is surprising just how much it dif-

[43] A scree plot demonstrated a rapid decrease in the explanatory value of the principal components. The first three principal components account for a significant portion of the variance and thereafter rapidly taper off to below 3 percent. See supplemental figures for this scree plot and to see this same data plotted along PC1/PC3 and PC2/PC3.

[44] *Zhi* can have a variety of meanings beyond its use as a particle. Other uses include as a pronoun and a verb meaning "to go."

ferentiates sections from historical works from novels, as novels, like historical documents, often situate events in clear geographical and temporal locations. [45] Documents that contain a preponderance of negations (不 *no*), informal language, and informal personal pronouns (你 *you*, *me*, *him*) spread from the bottom center to the middle right of the principal component space. These works are entirely novels.

Principal Component Loadings Plot
Official Histories and Novels
Ten thousand 1-gram sections, one thousand most common tokens

Figure 5B

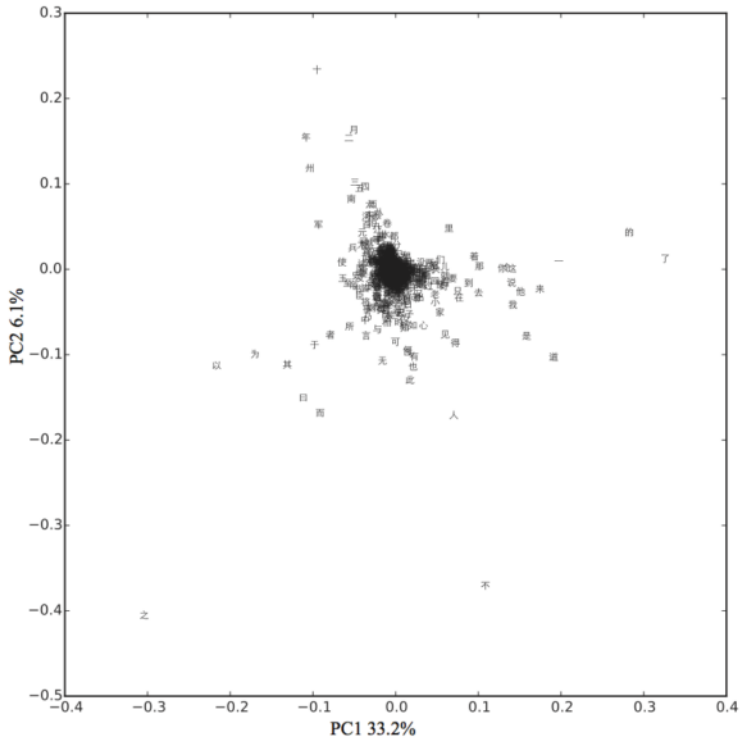


Figure 5B. Principal component loadings from Figure 5A. This loadings plot shows the influence the variables have on the principal components. It is clear that the first principal component is related to how classical or vernacular the prose in the document is. Characters on the left are more often used in classical works, while the characters on the right tend to dominate in vernacular speech. The second principal component is less clear but still shows some interesting behavior. The positive loadings clearly relate to geographical and temporal space. The negative loadings seem to be more closely related to people and feelings.

There are interesting interactions occurring along a diagonal formed by these principal components, but a detailed look at loadings across the individual principal components offer some more concrete conclusions. The first principal component shows a gradient from classical to vernacular language.

Table 1 shows the top negative and positive loadings along the first PC: [46]

Loading	Meaning	Score
之	possessive particle (classical)	-0.305
以	to use	-0.217
为	to make/to act as	-0.171

[45] The sections of histories in the upper left corner of this figure are mostly written in an annalistic format with running dates, which probably contributes to the sharp division.

[46] These definitions should be taken loosely. In classical Chinese in particular, characters can have a wide variety of meanings. And recall that these are 1-grams and not words.

Loading	Meaning	Score
之…之	possessive pronoun (classical)	-0.132
曰	to speak (classical)	-0.113
年	year	-0.109
州	province	-0.105
于	in	-0.099
十	ten	-0.096
军	army	-0.094
而	there	-0.092
之…	nominalizer	-0.080
使	to cause, to dispatch	-0.065
王	king	-0.064
至	arrive	-0.059
二	two	-0.057
所	location/object nominalizer	-0.056
之–	south	-0.054
之…之	troops	-0.053
月	month	-0.052
…	…	…
之–	to acquire	0,0718
只	only	0.074
在	at	0.076
到	arrive	0.088
着	particle, continued action	0.094
去	to go	0.100
那	that	0.100
不	no/negation	0.108
你	you	0.128
个	measure word	0.134
说	to speak (vernacular)	0.140
这	this	0.140
之’	me	0.141
之-	he	0.151
是	is (copula)/true	0.158
来	to come	0.174
之”	road	0.190
一	one	0.197
的	possessive particle (vernacular)	0.281
了	aspect particle	0.324

Table 1. Top twenty negative and positive loadings along PC1. These illustrate a strong classical to vernacular polarity along this principal component.

Note the several cases where the characters are cognates and essentially have the same meaning. The major difference is that those on the left tend to appear in works written in formal/classical Chinese, while those on the right tend to be more common in informal/vernacular works.

In the case of the second principal component, which explains around six percent of the variance, the negative loadings are a little less easy to interpret. 之 *possessive* and 不 *no* still explain much of the variance along this PC. Some of the less influential negative loadings could be interpreted to relate to people and interiority, but this may be reading too much into them. This ambiguity is not too unexpected, as the novels are fairly evenly distributed along the lower part of this PC. The positive loadings, however, are much more informative and explain much of the variation within historical texts

and are closely related to historicity in both time and space. Table 2 shows the top 20 positive loadings in PC2:

Loading	Meaning	Score
十	ten	0.233
月	month	0.163
年	year	0.154
二	two	0.154
州	province	0.118
三	three	0.102
四	four	0.097
☒”	five	0,09
☒–	south	0.082
西	west	0.076
☒…☒	six	0.076
☒…☒	eight	0.067
县	county	0.064
东	east	0.063
七	seven	0.062
河	river	0.062
☒–	north	0.054
卷	scroll	0.054
百	one hundred	0.052
军	army	0.052

Table 2. Top 20 positive loadings along PC2: These loadings are strongly associated with numbers and geographical space.

The documents that fall on the upper extremity of the second principal components are the annalistic portions of the official histories, while the more prose-driven sections fall nearer to the novels.

The intermingling of several novels into official history space in the lower left quadrant of Figure 5A suggests that, in some subsections at least, novels and official histories can share similar styles. I would hypothesize these texts are mostly novels on historical topics. This intermingling is less clear when looking at a PCA of the full texts. Although they distinctly cluster, as shown in Figure 6, the nature of the space between them is vague.

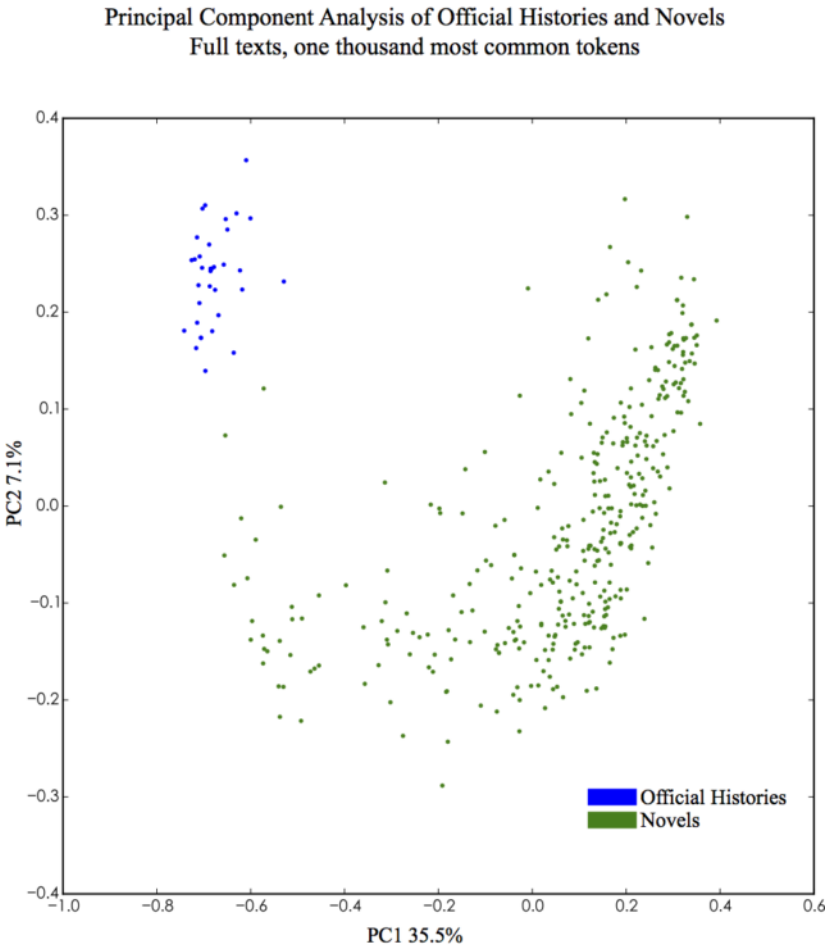


Figure 6. Principal component analysis of the full texts of official histories and novels using the one thousand most common tokens. The shape of the PC space is preserved from Figure 8 but there is less clear mingling of novels in history space. This is not too surprising and suggests that while some interior sections bear distinct similarities, these similarities are obscured when looking at full texts. The second PC is no longer dominated by numbers. Instead, it is heavily influenced in the positive direction by words with governmental connotations. The negative direction is vaguely associated with people and emotion.

The second PC in Figure 6 is more interesting than in the last example, offering an interesting scale from interiority (知 *to know*, 心 *heart*, 见 *to see*) in the negative loadings to some historicity in the positive loadings (军 *army*, 皇 and 帝 *emperor*, 王 *king*, 州 *province*). The influence of the number-heavy annalistic sections is much reduced.

To further parse the relationships among these texts, and to understand their stylistics, it makes sense to integrate historical romances. These offer an interesting stylistic and content-driven compromise between purely fictional novels and histories. Historical romances were fictional retellings of historical events, so in content they are similar to histories, and in terms of rhetoric were similar to fictional works. Though many were written in a simplified classical form, many others were purely vernacular. In a traditional sense, dividing novels and *yanyi* is a bit odd: most bibliographers consider *yanyi* to be a form of *xiaoshuo*. The categorization schema of the Wenxian website provides an easy to use division that facilitates this analysis.

Figure 7 shows novels, historical romances, and official histories. The shape of this space is similar to Figure 5A, but the historical romances fit very neatly in the margins between fiction and official histories. They bleed extensively into the novels and marginally into the histories. These texts appear to fall along a multi-dimensional stylistic gradient that emerges in this principal component space with poles in vernacular, classical, historical, and fictional language.

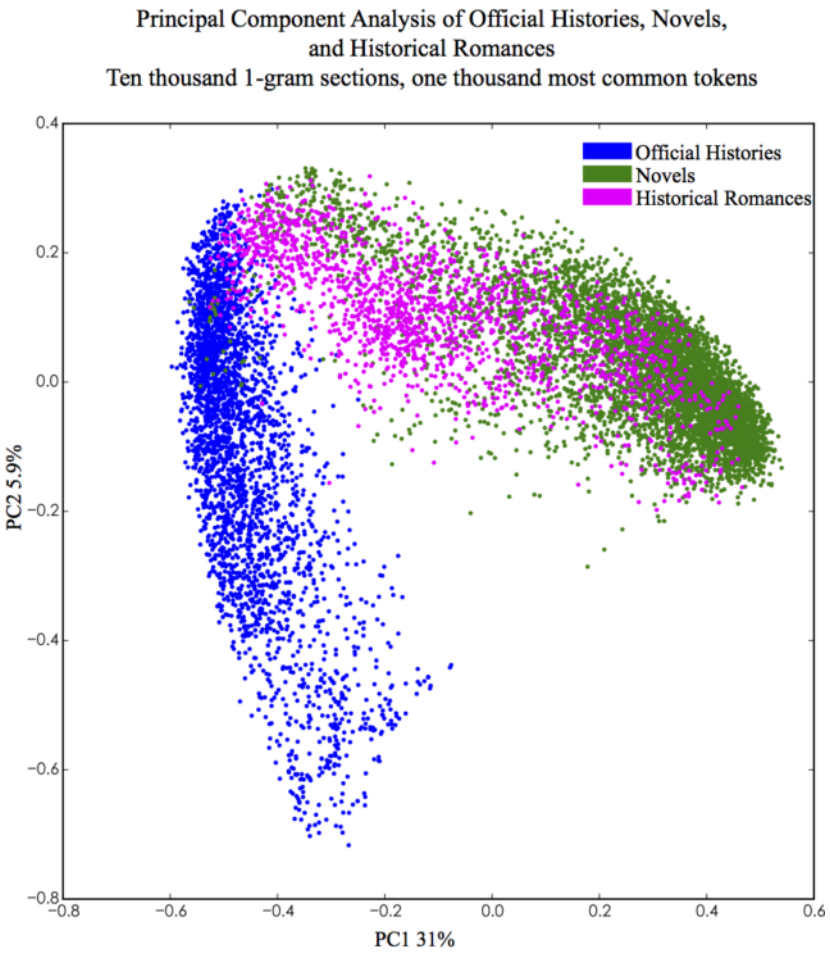


Figure 7. Principal component analysis of ten thousand 1-gram sections from official histories, novels, and historical romances. When historical romances are added to the picture, the general shape of the PC space remains the same but is flipped. The historical romances lay along the same space that novels occupy, but intrude further into the space dominated by the histories. The same linear cleaving between the more fictional and historical works is still evident.

Figure 8 shows the principal component space of these works when all *yeshi* identified in the Beijing collection are added. Most documents from *yeshi* fall in space dominated by official histories and lie transversely along documents from official histories. Yet a significant portion extends into space occupied by the historical romances, with some mingling among the pure novels. They form an apparent bridge between the novelistic and historical genres.

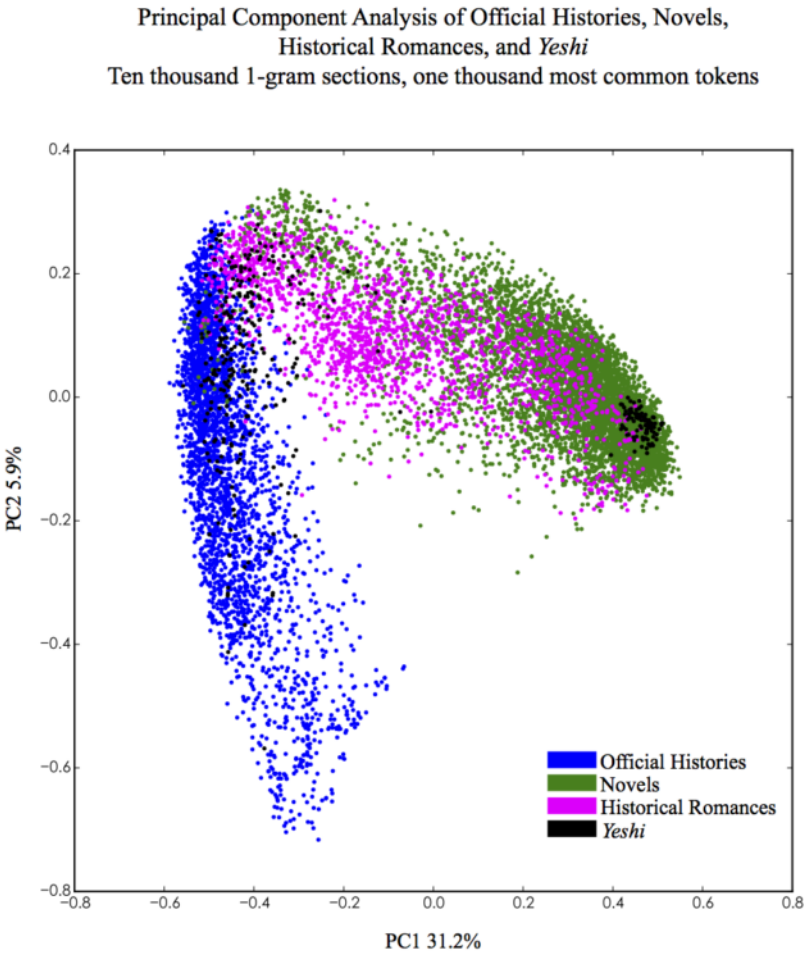


Figure 8. Principal component analysis of official histories, novels, historical romances, and *yeshi*. When *yeshi* enter in to the picture, they share a great deal of similarity to the official histories, overlap into space mostly dominated by historical romances, and even, in two cases, enter into dense novel space on the right. This is particularly interesting: The two *yeshi* are actually republican era novels masquerading as *yeshi*.

The ten thousand character sections that fall in the densest part of novel space on the middle right of Figure 8 illustrate the complex nature of *yeshi*. These sections come from two texts, *Outer History of Eastern Liu* and *Outer History of Eastern Liu, continued*. These are early Republican era texts that are generally regarded as novels. In fact, according to their original classification on the Wenxian website, they are novels. [47] Figure 8 amply demonstrates that *yeshi* are distributed on the plot in a manner that accords with their complex stylistic, bibliographic, and cultural nature. *Yeshi*, when taken in consideration with historical romances, bridge the stylistic gap between pure fiction on the one hand and pure history on the other.

There are several limitations imposed by the corpus used above. Many texts fall outside the late Imperial period of interest and there are not too many *yeshi*. There is an alternative corpus available that can represent *yeshi*. The Wenxian website contains a category of text in the historical section called the *Stored Records and Recorded Annals* that comprises a large number of unofficial histories, among a few other types of text. [48] This corpus, limited to texts written during the Ming and Qing periods is included here in Figure 9. These new texts share significant space with official histories and formal historical romances, but also extend into vernacular novel space. The overall shape of the PC space illustrates the bridging effect that the *yeshi* have in the stylistic space formed by these texts.

[47] Recall, the *yeshi* label is imposed from outside the corpus.

[48] Preliminary research suggests that many of these works are *yeshi*, and those that aren't appear to be stylistically similar. I showed the initial analysis with fewer *yeshi* to illustrate the similarities between the corpora. This corpus adds an additional 508 texts, which split into 1,620 sections.

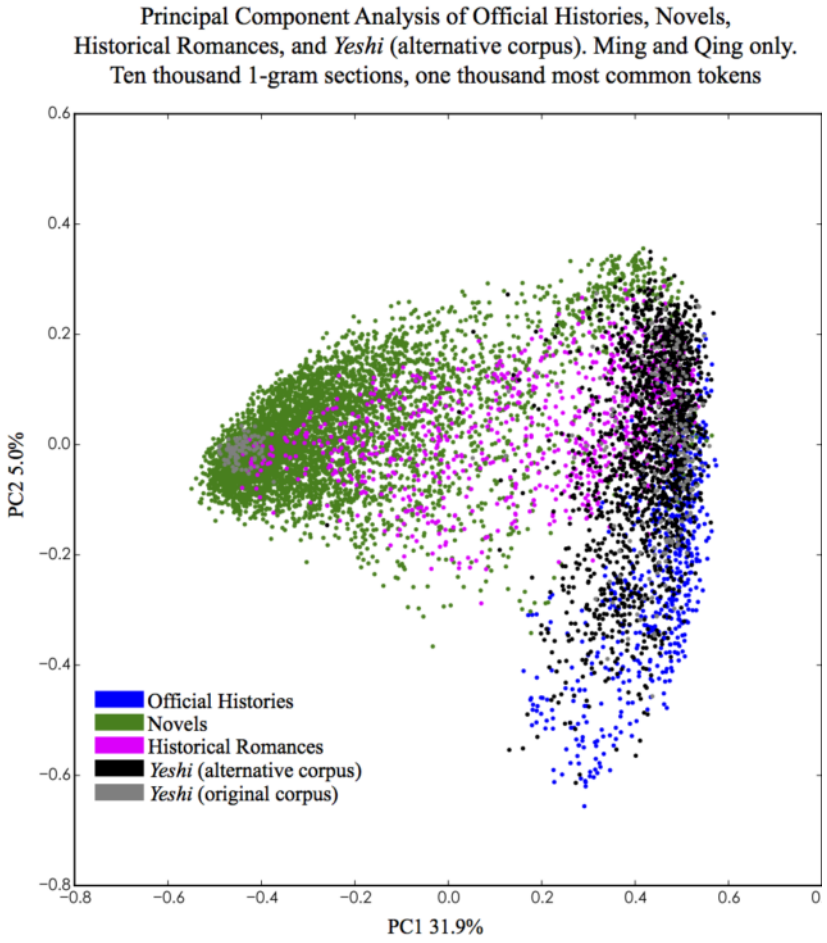


Figure 9. Principal component analysis of official histories, novels, historical romances, and texts taken from wenxian.fanren8.com's Stored Records and Recorded Annals collection, which contains a significant number of *yeshi*, among other texts. This corpus is limited to texts written in the Ming and the Qing dynasties. The shape is similar to previous figures, but the overlap between *yeshi* and official histories is even more distinct, with significant intrusions into fictional space.

Loadings prove an effective way of understanding the linguistic relationships among these texts, but there are other ways of determining which characters are most effective in differentiating these texts. A linear support vector machine used to select important features from the corpus used in Figure 8 provides some insight, in spite of the significant overlap in 1-gram usage among some of these works. It finds that the 20 most important features in differentiating these works are 一 *one*, 不 *no*, 为 *to make/to act as*, 之 *possessive*, 了 aspect particle, 人 *person*, 以 *to use*, 你 *you*, *troops*, 军 *army*, 十 *ten*, 后 *after*, 州 *province*, 年 *year*, *me/I*, 来 *to come*, 王 *king*, *to be born/Mr.*, 说 *to speak*, *road*. These features are strikingly similar to the loadings across the first two principal components seen earlier. This reinforces the conclusions drawn earlier, that a spectrum of formality and historicity form the structure of the differences among these texts. This also shows that because the classical and vernacular variant of a specific term are very negatively correlated, the opposite cognate is not generally needed for differentiation.

Conclusions

Both hierarchical cluster analysis and principal component analysis offer insight into the stylistic relationships between fictional and historical narrative in imperial Chinese literature. Hierarchical cluster analysis illustrates the ease with which texts in Chinese genres cluster, and hints toward a broader linguistic polarity defined by classical and vernacular linguistic

registers. This is partially an artifact of the chosen corpus. If it contained entirely poetic works, this dichotomy would likely not emerge. Yet it does when looking at a broad spectrum of fictional and historical prose.

Principal component analysis breaks the polarity seen in the HCA down and reveals a multi-dimensional stylistic gradient evident in the first two principal components across a number of different corpora. Official histories connect to novels through intermediate *yeshi* and historical romances. [49] This continuum is largely delineated by the nature of the prose in the texts: *yeshi* and official histories in general are dominated by classical language, while the novels show a wide linguistic register, with a significant portion written in very vernacular language. Within the historical genres there is another range of differentiation that lies along a historical/interiority gradient. The official histories are dominated by words of strong historical and geographic import, while the novels are lightly influenced by negations, people, and limited words that imply some interiority.

Some of these conclusions are unsurprising: we have long known that fiction has existed on a linguistic continuum between vernacular and classical, official histories are written in very formal language, and *yeshi* are fairly formal texts. What we can now do is place specific texts within this continuum. Furthermore, quantitative analysis allows us to easily identify and track down bibliographic exceptions, highlighting them for closer inspection.

We have confirmation that *yeshi* is a semi-variable genre that stylistically mingles with official works, while also extending into linguistic territory dominated by novels. [50] This speaks directly to late Imperial and modern discussions surrounding *yeshi* as a genre of historical text marginalized in official discourse. *Yeshi*'s homology with official historical works and occasional extension into novelistic space is clearly evident, underscoring their unstable position. *Yeshi* that fall squarely within novelistic space reveal important information about what constitutes a *yeshi*, even if it is the understanding that the term *yeshi* is an unstable construct that can variably be a misclassification or a conscious choice made by a novel author to frame their story in a wider historical/fictional discourse.

Most critically, quantitative analysis reveals that historical content is marginally predictive but not determinative of style. Recall that in Figure 3, texts on Wei Zhongxian that contain nearly identical historical content cluster on different sides of the figure, depending on genre. Yet this apparent dichotomy between *yeshi* and vernacular novels breaks down when studying large systems of texts. What seems obvious in individual cases is not an accurate picture of macro-level features. Interestingly, the stylistic influence of history holds true even when words with strong historical semantic meaning are limited in the analysis. [51] The PCA loadings suggest that semantically important historical terms play a role in differentiating the texts here, but other more basic words are just as important. This implies that historical discourse exerts a structural influence on the use of language within a text at a very basic level.

Paul Vierthaler, Boston College

Appendix 1: Word Segmentation in Chinese

Algorithmically parsing word boundaries in Chinese texts is an area of active research among computer scientists and linguists. Although the primary goal of most scholars working on the problem is focused on machine translation and text-to-speech software, the implications are important for literary scholars. [52] Although systems are become increasingly sophisticated, few algorithms are capable of achieving higher than 95 percent accuracy. [53]

As early as 1990, statisticians and linguists were already developing probabilistic methods for dividing Chinese *ci*. In 1990, Richard Sproat and Chilin Shi were able to achieve decent rates of boundary recognition by looking at the frequency of *zi*

[49] I also tested texts of genres that one would not expect to find on this gradient. They tend to fall either to the side of the gradient or in space that is near 0 on the second principal component, but to the more classical side of the first principal component, situating them in the space where these texts intermingle the most. They don't usually bleed much into the other genres in the way that the texts under investigation do.

[50] Like most analyses of this type, an expanded corpus with more detailed metadata would significantly increase the rigor of the findings. Exact feature selection would also offer greater insight into how to differentiate historical and fictional works.

[51] An interesting avenue of future research might be to completely eliminate these terms and see if we can predict historical content.

[52] Richard Sproat and Chilin Shih, "A Statistical Method for Finding Word Boundaries in Chinese Text," *Computer Processing of Chinese & Oriental Languages*, 4.4 (1990), 337. Shoushan Li and Chu-Ren Huang, "Word Boundary Decision with CRF for Chinese Word Segmentation," *23rd Pacific Asia Conference on Language, Information and Computation* (2009), 726. Gaoqi Rao and Endong Xun,"Word Boundary Information and Chinese Word Segmentation," *International Journal on Asian Language Processing*, 22.1 (2012), 16.

[53] The upper threshold for many approaches is around 90 to 95 percent. For example: Sproat, "A Statistical Method," 345.

associations throughout a corpus of Chinese newspapers. [54] They then grouped the *zi* into *ci* depending on the strongest positive correlation. In Sproat's tests they were able to group approximately 90 percent of *zi* correctly. [55]

Since Sproat's work was first published there have been numerous advancements in word segmentation. Most recent methods use supervised machine learning to procedurally generate rules to more accurately find word boundaries. [56] Others use "maximum matching" algorithms that depend on matching *ci* in a text with an existing lexicon guided by hand-written rules. [57] It seems that these machine learning-based algorithms show the most promise, but they often require extensive training sets using pre-parsed texts. There have been some recent developments in word segmentation technology that do not require a training set. These unsupervised algorithms show promise for those interested in studying literature written in less-studied periods. [58]

Some of these methods are becoming accessible to average scholars. The Stanford Word Segmenter is a publicly available machine learning segmenter that takes into consideration both lexical features and the context in which *ci* appear. [59] It can use one of two training sets: the Penn Chinese Treebank (CTB) and the Peking University Standard (PKU). The Penn Chinese Treebank is the larger of the two training sets, is manually segmented, and contains parts-of-speech markers. [60] The Stanford word parser works for modern Chinese [61] but is only marginally accurate in parsing classical Chinese.

There are advantages and disadvantages to using character *n*-grams as tokens, as opposed to *ci* "words" or word *n*-grams. The main advantage of using *n*-gram tokenization is there is a determinative result, while it is not immediately obvious how accurately computerized algorithms parse classical Chinese texts. The lack of a sharp distinction in classical Chinese between *zi* and *ci* [62] means that *n*-gram analysis largely obviates the problems introduced by a lack of good classical Chinese parsers. However, 1-gram based analysis often breaks the language down further than is justified. Even in classical Chinese, in which most words are single characters, compound-character words exist. [63] Two and three-gram analysis add information on context, but as *n* increases sparse data rapidly becomes an issue. [64]

[54] Sproat, "A Statistical Method," 339.

[55] Sproat, "A Statistical Method," 343.

[56] Machine learning is a type of artificial intelligence in which a computer program trains on a data set with known characteristics and then generalizes from the training set to classify data outside the training set. For example, Ka Seng Leong *et al.*, "Chinese Word Boundaries Detection Based on Maximum Entropy Model" (paper presented APCOM '07, Kyoto, Japan, December 3-6, 2007). Peng Fuchun, Fangfang Feng, and Andrew McCallum, "Chinese Segmentation and new Word Detection using Conditional Random Fields," *Proceedings of the 20th Annual International Conference on Computational Linguistics* (2004), article 562.

[57] Chih-Hao Tsai, "MMSEG: A word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm," accessed March 26, 2014.

[58] This research has not been published yet but represents an important advancement in Chinese word segmentation. Ke Deng, Bol P.K., Kate J. Li and Liu J.S. (2015+) word segmentation. Ke Deng, Bol P.K., Kate J. Li and Liu J.S. (2015+). On Unsupervised Chinese Text Mining. Invited Revision by PNAS.

[59] It uses the Conditional Random Field described in Huihsin Tseng et al., "A Conditional Random Field Word Segmenter," *Fourth SIGHAN Workshop on Chinese Language*, (2005). They have additionally added a lexical element similar to Pi-Chuan Chang, Michel Galley and Chris Manning, "Optimizing Chinese Word Segmentation for Machine Translation Performance," *WMT* (2008), accessed March 24, 2014, http://nlp.stanford.edu/pubs/acl-wmt08-cws.pdf. Software is available at http://nlp.stanford.edu/software/segmenter.shtml. I used version 3.3.1 when dividing the texts into *ci*.

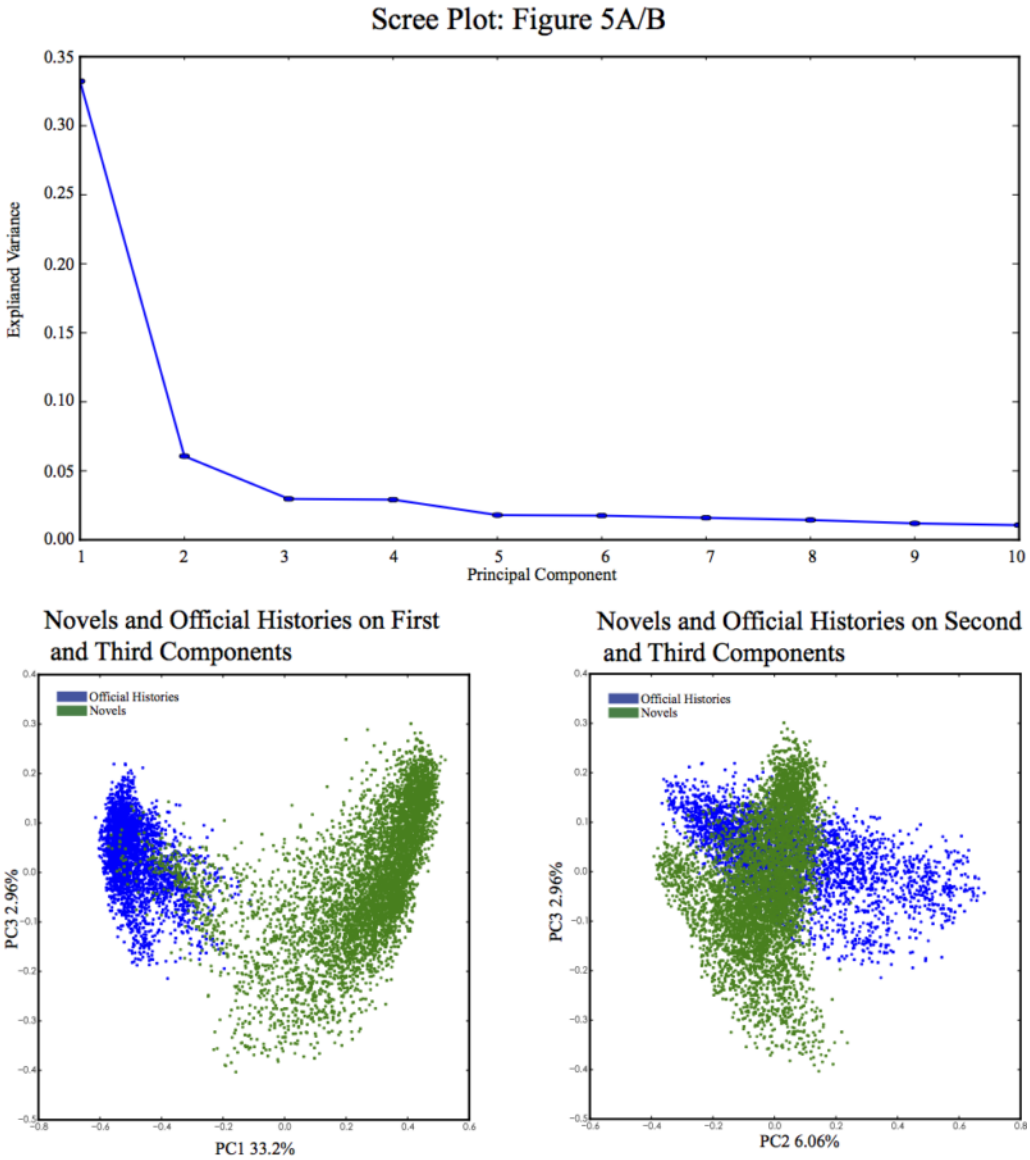
[60] Nianwen Xue, Fu-Dong Chiou, and Martha Palmer, "Building a Large Annotated Chinese Corpus," *Proceedings of the 19th International Conference on Computational Linguistics* (2002), Taipei, Taiwan. Accessed on March 26, 2014, available at http://www.cis.upenn.edu/~chinese/coling02.ps.

[61] Tseng, "A Conditional Random Field Word Segmenter," 3.

[62] Personal conversation, Guo Yingde, April 2014.

[63] Xue Nianwen uses the example of the *zi chan* 自产 to produce. It can occur in multiple places within a *ci*, or as a *ci* on its own. Nianwen Xue, "Chinese Word Segmentation as Character Tagging," *Computation Linguistics and Chinese Language Processing* 8.1 (2003): 30.

[64] Even 4 or 5-grams create very sparse datasets in classical Chinese.



Supplementary Figure 1 Scree Plot: This plot shows the explained variance of the top ten components in the dataset used for Figure 8 and 9.
Supplementary Figure 1 PC1 and PC3: PC3 only explains about three percent of the variance within the dataset and does not offer much in the way of explanation.
Supplementary Figure 1 PC2 and PC3: PC2 and PC3 together do not offer an excellent explanation of the relationship among these texts.